Variance Estimation for the Multidisciplinary Treatment Planning (MTP) Survey

Hyunshik J. Lee[1], Benmei Liu[2], Irene Prabhu Das[3] Mary Fennell[4]
[1]Westat, 1600 Research Boulevard, Rockville, MD 20850
[2]National Cancer Institute, 6116 Executive Boulevard, Suite 504, Rockville, MD 20852
[3] National Cancer Institute, EPN 4095B, 6130 Executive Boulevard,
Bethesda, MD 20892
[4]Brown University, Box 1916, Providence, RI 02912

Abstract
The Multidisciplinary Treatment Planning (MTP) Survey will be conducted using an instrument jointly developed by the National Cancer Institute (NCI) and the Commission on Cancer (CoC) to survey CoC-accredited hospitals in the United States on how multidisciplinary treatment planning is offered to cancer patients for one of six cancer types (Brain/CNS, Lung, Head & Neck, Gynecologic, Gastrointestinal, and Breast cancer). This survey used a census of facilities and only one cancer type is randomly selected for each facility by an unequal probability sampling method using case volume as the basis to determine the selection probability. The sampling method poses a unique challenge for variance estimation because a single cancer type is selected independently from each facility, which can be regarded as the sampling stratum because each facility is included with certainty. It requires combining facilities into variance strata and special handling of nonresponse. This paper presents a theoretical framework for variance estimation and the results of a simulation study of the proposed variance estimator based on the jackknife technique.

Key Words: Jackknife variance estimator, variance stratum/unit, combining strata

1. Introduction

The main purpose of the survey is to understand the various ways in which multidisciplinary care is defined, structured, and implemented for different types of cancers in facilities that have cancer programs accredited by the American College of Surgeons Commission on Cancer (CoC). Six cancer types will be considered in this study, Brain/CNS (BCNS), Lung, Head & Neck (HN), Gynecologic (GYN), Gastrointestinal (GI), and Breast cancer. Only one of the six cancer types will be addressed in the survey for each hospital.

The Cancer Liaison Physicians (CLP) at the CoC-accredited hospitals will constitute the survey respondents/participants in this study. There are 1,309 facilities (as of August 2011) after removing the following facilities to be excluded from the survey: a) NCI Community Cancer Centers Program 2007 and 2010 sites, and b) programs from which CLPs had participated in the cognitive testing. All the CoC-accredited facilities with the exception of a small number of facilities that are designated to be excluded will be invited to participate in the study. The CLPs at each facility will be asked to complete the

survey about MTP at their respective hospital for just one cancer type sampled/selected from up to six cancer types with sufficient patient case volume.

The sampling frame provided to the National Cancer Institute by the CoC includes a facility ID number and a designation of case volume for each of the six cancer types categorized into three groups: fewer than 12 patient cases per year (coded as 0), at least 12 patient cases per year or less than the average number of cases across all programs that had at least 12 patient cases per year (coded as 1), and the average or higher than the average across all programs that had at least 12 patient cases per year (coded as 2).

Actual case volumes are not provided due to confidentiality reasons. The case volume category of 0 for a particular cancer type is considered too small and therefore, cancer programs with category 0 volume for such cancer types would not be assigned that cancer type survey. The sampling frame of 1,309 facilities includes 17 facilities with a 0 case volume code for all six cancer types. Thus, these 17 facilities will be deleted from the sampling frame. The final census of eligible CoC-accredited hospitals (i.e., they have at least one cancer type with case volume coded as 1 or 2) for the survey is 1,292. Hospitals that have either case volume codes of 1 or 2 for each of the six cancer types are eligible.

To assign the cancer type to the hospital and facilitate meaningful data analysis by providing a means of estimating the standard error, a probability sampling method is applied to each facility in order to select the cancer type for which the CLP needs to answer for their facility. The probability sampling method is described in more detail in the next section.

The table in Appendix A provides the distribution of the 1,309 facilities by cancer type and case volume. As shown in the table, there are 121 unique combinations (labeled as "group number") of cancer type by case volume of the 1,292 facilities with at least one (eligible) cancer type (i.e., 17 ineligible facilities are excluded).

2. Sample Design

All the 1,292 facilities with at least one cancer type with case volume codes of 1 or 2 will be asked to participate in the survey but each facility will be asked to complete just one questionnaire for only one selected cancer type. To accomplish the selection of one cancer type for each facility, a probability sampling method is used.

Recognizing the importance of cancer types with higher case volume, it was decided to select a cancer type with case volume code (CVC) = 2 with a larger probability than a cancer type with CVC = 1. The cancer type selection probability is determined by the following formula:

$$p_{ij} = v_{ij} / \sum_j v_{ij} \tag{2.1}$$

where v_{ij} is the case volume code (i.e., 0, 1 or 2) for facility i, cancer type j. Note that if $v_{ij} = 0$, then the probability is zero, and such cancer type is never selected. Also note that if v_{ij} are all equal within a facility, then the probabilities are all equal, and selection of a cancer type is random across different cancer types. This means that each cancer type within the facility has an equal chance of being selected. If a facility has only one eligible cancer type, then its probability is one, and the cancer type is selected with certainty.

Within facilities having cancer types with mixed case volume codes, the cancer types with CVC = 2 are given a probability twice as large as those with CVC = 1.

Following the above sampling scheme, the total number of facilities to be assigned for a certain cancer type is random. Table 1 provides the expected sample sizes along with the population size by cancer type. The sum of the expected sample sizes is equal to 1,292. The table also provides the standard deviation of the sample size, which is useful to see how variable the sample size is.

Table 1: Population and Sample Sizes by Cancer Type

	Cancer type						
	BCNS	*HN*	*GYN*	*Lung*	*GI*	*Breast*	*Total*
Population size: CVC = 1	547	513	735	796	871	833	4,295
Population size: CVC = 2	220	192	327	466	408	440	2,053
Population size: total	767	705	1,062	1,262	1,279	1,273	6,348
Expected sample size	129.5	110.0	202.9	282.2	282.7	284.6	1,292
Standard deviation	10.0	9.6	12.7	14.6	14.6	14.6	

The data analysis will be focused on the full sample rather than by cancer type. Thus the randomness of the sample size for cancer types does not matter.

3. Variance Estimation

For a given survey characteristic y, its value for facility i, cancer type j is denoted by y_{ij}. Let Y be the total of the variable for the whole population, that is, $Y = \sum_{i=1}^{1292} \sum_j y_{ij}$. For the time being, we assume that we are interested in estimating the total Y. Totals are of interest because they are basis of many parameters and corresponding statistics (e.g., means and proportions), and the usual total estimate of the weighted sum is a linear statistic, whose variance can be more easily derived.

From each facility, one cancer type is selected with a probability p_{ij} as defined by (2.1). Let w_i be the weight associated with the selected cancer type as defined by $1/p_{ij'}$, where j' is the cancer type selected. Then Y is estimated without bias by

$$\hat{Y} = \sum_{i=1}^{1292} w_i y_i \tag{3.1}$$

where $y_i = y_{ij'}$, which is the y-value of the selected cancer type j' from facility i. Because the sampling is done independently for each facility, each facility is a stratum, and the variance of the estimator (3.1) can be written as:

$$V(\hat{Y}) = \sum_{i=1}^{1292} V(w_i y_i) \tag{3.2}$$

However, this variance cannot be estimated because $V(w_i y_i)$ cannot be estimated for the reason that only one cancer type is selected from each facility. It is easy to see that

$$V(w_i y_i) = E\{w_i y_i - E(w_i y_i)\}^2 = \sum_j w_{ij} y_{ij}^2 - Y_i^2 \tag{3.3}$$

where $Y_i = \sum_j y_{ij}$ and the summation is over cancer types with non-zero case volume

code. Note that this is a theoretical population value, which cannot be estimated from a sample.

Wolter (2007) discusses ways of handling this situation. The main idea is to collapse strata into a smaller number of variance strata so that each new (variance) stratum has more than one sample unit; the method is called the collapsed stratum estimator (Wolter, 2007, p. 50). The collapsed strata are called the variance strata because they are formed for variance estimation. We can define many different collapsed variance estimators depending on how to collapse strata. However, the most basic one is to pair strata to form variance strata with two sample units in each. Then the collapsed variance estimator is given by:

$$v_{\text{cs}}(\hat{Y}) = \sum_{h=1}^{646} v_{\text{cs}}(w_{h1}y_{h1} + w_{h2}y_{h2}) = \sum_{h=1}^{646}(w_{h1}y_{h1} - w_{h2}y_{h2})^2 \qquad (3.4)$$

where w_{hi} and y_{hi} are, respectively, the sampling weight and y-value of facility i ($i = 1$ or 2) in paired stratum h. The bias of this collapsed variance estimator is:

$$B\{v_{\text{cs}}(\hat{Y})\} = \sum_{h=1}^{646}(Y_{h1} - Y_{h2})^2 \qquad (3.5)$$

where Y_{hi} is the stratum total for facility i ($i = 1$ or 2) in paired stratum h. Therefore, if $Y_{h1} - Y_{h2} = 0$ for all paired variance strata, then the collapsed variance estimator is unbiased. Expression (3.5) provides a strategy for how to pair the facilities to minimize the bias. Assuming that the cancer type structure and case volume are predictive of the facility total Y_{hi}, we pair the facilities according to their similarity in the pattern of cancer type and case volume. As shown in Appendix A, there are 121 unique patterns of cancer type and case volume. Assuming that the *Y*-total at each facility is similar for the facilities with the same pattern of cancer type and case volume, we pair facilities within each of the 121 unique patterns. If there are more than 2 facilities, pairing is done randomly. The paired facilities form the collapsed strata. With such pairing, the bias given in (3.5) should be small. One issue we have to deal with in this pairing exercise is the unique pattern groups with an odd number of facilities, or just one facility (there are a few groups with one facility). If a unique pattern group has only one facility, then it is collapsed into the adjacent group. If the number of facilities in a unique pattern group is odd and greater than 2, one of the collapsed strata contains triple facilities. There are still over 600 collapsed strata after this pairing strategy.

Another strategy is to collapse facilities with similar cancer types and case volume codes into larger groups. We call this collapsing strategy as Strategy 1, and the pairing strategy described above as Strategy 2. These two strategies are summarized as follows:

- Strategy 1: Group facilities based on the similarity of cancer type and CVC composition, and define collapsed strata by the grouping.
- Strategy 2: Group facilities with exactly the same cancer type and CVC structure, and define collapsed strata by random pairs of facilities within the groups.

The idea is to group similar facilities assuming that the variable values are similar within such groups. To form Strategy 1 grouping, we use the following four criteria.

1. The number of eligible cancer types is the same;

2. The sum of case volume numbers is the same or similar;
3. There should be at least 12 facilities in each stratum but preferably 18 or more;
4. Form a different stratum consisting of facilities with the same case volume distribution pattern as long as criterion 3 is satisfied.

The variance strata formed in this way are shown in Table 2. The Eligible Count column gives the number of eligible cancer types, and its value ranges from 1 to 6. This is a primary criterion to form a (collapsed) variance stratum. For example, Variance Stratum 1 consists of facilities with only one eligible cancer type. The Sum of Case Volume Codes column provides the facility's sum of case volume code values. For example, if a facility has two eligible cancer types, Breast with CVC = 1 and GYN with CVC = 2, its Sum of Case Volume Codes is 3. The Count column is for the count of facilities for each unique combination of Eligible Count and Sum of Case Volume, and the Total Count column presents the size of Variance Stratum, which may consist of more than one unique combination. The No. of random groups column shows the number of random groups to be formed.

Table 2: Definition of Variance Strata, their Size and Number of Random Groups

Variance stratum	*Eligible number of cancer types*	*Sum of case volume codes*	*Count*	*Total count*	*No. of random groups*
1	1	1	14	14	2
2	2	2	16	18	3
		3	1		
		4	1		
3	3	3	147	149	25
		4	1		
		5	1		
4	4	4	19	19	3
5	4	4	196	196	33
6	4	4	19	19	3
7	4	5	25	25	4
8	4	6	11	13	2
		7	2		
9	5	5	83	83	14
10	5	5	84	84	14
11	5	6	22	22	4
12	5	6	18	18	3
13	5	6	18	18	3
14	5	7	20	20	3
15	5	8	13	13	2
16	5	9	10	13	2
		10	3		
17	6	6	86	86	14
18	6	7	19	19	3
19	6	7	24	24	4
20	6	7	39	39	6
21	6	8	24	24	4
22	6	8	23	23	4

Table 2: Definition of Variance Strata, their Size and Number of Random Groups (Continued)

Variance stratum	*Eligible number of cancer types*	*Sum of case volume codes*	*Count*	*Total count*	*No. of random groups*
23	6	9	19	19	3
24	6	9	45	45	7
25	6	9	18	18	3
26	6	10	31	31	5
27	6	10	48	48	8
28	6	11	45	45	7
29	6	11	43	43	7
30	6	12	104	104	17
Total			1,292	1,292	212

The fourth criterion is to ensure a large number of facilities to calculate the variance within variance stratum for variance estimation. Variance Strata 4, 5, and 6 in the table look the same according to the combination of Eligible Count and Sum of Case Volume Codes but the case volume distribution pattern is different (see Group Numbers 14, 16, and 55 for Variance Strata 4, 5, and 6, respectively in the table in Appendix A). Since each pattern in this situation has a large enough number of facilities, a separate variance stratum is created. Note that Variance Stratum 1 is an absolute certainty.

There are two main variance estimation methods: Taylor method and resampling method. For the survey, we intend to use the latter method for its well-known advantages, one of which is that it does not require the linearization as the Taylor method requires. Among different resampling variance estimators, we studied the following ones:

- Based on Strategy 1 variance strata, the full jackknife variance estimator (coded as JKn);
- Based on Strategy 1 variance strata and random groups (variance units), the jackknife variance estimator (coded as JKn_RG);
- Based on Strategy 2 paired variance strata, the paired jackknife variance estimator (coded as JK2);
- Based on Strategy 2 paired variance strata, the Rao-Wu rescaling Bootstrap methods (Rao, Wu, and Yue, 1992) with three bootstrap sample sizes, $B = 100, 200,$ or 500.

The full jackknife variance estimator with the maximum number of replicates, which is defined by dropping one facility at a time to create a replicate called the JKn, has 1,278 replicates after removing 14 absolute certainties. In this case, the single facility is the variance unit (the base unit for variance estimation), and the resulting variance estimator will be more stable with the largest possible degrees of freedom. However, it is cumbersome to carry so many replicate weights, and more computing time will be needed for analysis. To reduce the number of replicates, we use random groups of facilities within variance strata to define the variance units. The last column of table 2 shows the number of random groups to be formed. The total number of random groups is 212 but there will be 210 replicates after removing absolute certainty random groups. Variance estimation formulae for the three resampling methods are provided in Appendix B.

4. Simulation Study

A simulation study was conducted to see how well the variance estimators proposed in section 3 work under some hypothesized populations.

For simulation, 16 variables, 8 continuous and 8 binomial, were generated as follows:

$$x_{ij} = \begin{cases} N(\mu_1, \sigma^2) & \text{if } c_{ij} = 1 \\ N(\mu_2, \sigma^2) & \text{if } c_{ij} = 2 \end{cases}$$
$$y_{ij} = \begin{cases} B(P_1) & \text{if } c_{ij} = 1 \\ B(P_2) & \text{if } c_{ij} = 2 \end{cases} \quad (4.1)$$

where c_{ij} is case volume code for cancer type j in facility i, $N(0,1)$ is the standard normal variate, μ and σ are mean and scale parameters, and $B(P)$ is a binomial variate with a parameter of P. The following tables show those parameters used to generate the variables and their population characteristics.

Table 3: Parameters used to Generate the Continuous Variables and their Population Characteristics

Continuous variable	*I/D*[1]	*Mean (μ)*		*Scale (σ)*	*Mean*	*Standard deviation*	*Correlation with CVC*
		CVC=1	*CVC=2*				
x1	I	20	20	3	20.00	3.02	0.016
x2	D	20	40	3	26.47	9.88	0.952
x3	I	50	50	3	49.94	3.00	0.024
x4	D	50	100	3	66.12	23.65	0.992
x5	I	50	50	6	49.89	5.99	0.024
x6	D	50	100	6	66.06	24.29	0.969
x7	I	30	30	6	30.08	6.01	-0.012
x8	D	30	60	6	39.78	15.20	0.918

[1] Note: "I" means that the variable is independent of CVC, and "D" means the opposite. The correlation column shows the strength of the relation between the variable and CVC.

Table 4: Parameters used to Generate the Binomial Variables and their Population Characteristics

Binomial variable	I/D[1]	*Proportion (P)*		*Mean*	*Standard deviation*	*Correlation with CVC*
		CVC=1	*CVC=2*			
y1	I	0.4	0.4	0.4145	0.4927	-0.008
y2	D	0.3	0.4	0.3467	0.4760	0.090
y3	I	0.3	0.3	0.3015	0.4590	-0.001
y4	D	0.2	0.4	0.2692	0.4436	0.198
y5	I	0.1	0.1	0.1038	0.3050	0.002
y6	D	0.05	0.2	0.1002	0.3003	0.229
y7	I	0.5	0.5	0.4989	0.5000	0.001
y8	D	0.6	0.3	0.5038	0.5000	-0.274

[1] Note: "I" means that the variable is independent of CVC, and "D" means the opposite. The correlation column shows the strength of the relation between the variable and CVC.

When a continuous variable used in the simulation is dependent with CVC, the correlation is high. However, this is not the case for binomial distributions.

The simulation results with 1,000 simulated samples are shown in the following tables in terms of the relative bias and relative standard error (i.e., coefficient of variation) of the variance estimators. The relative bias (Rel.Bias) and relative standard error (Rel.SE) of a variance estimator are defined as follows:

$$\text{Rel. Bias}(\hat{V}) = E(\hat{V})/V - 1 \quad (4.2)$$

$$\text{Rel. SE}(\hat{V}) = \sqrt{V(\hat{V})}/V \quad (4.3)$$

where $E(\hat{V})$ and $V(\hat{V})$ are the expectation and variance of a variance estimator $\hat{V}$. The bootstrap variance estimators are denoted in the tables as Bootstrap1, Bootstrap2, and Bootstrap3, respectively, for the bootstrap sample sizes of 100, 200, and 500.

Table 5: Simulation Results – Relative Bias of the Variance Estimator

Variable	*JKn*	*JKn_RG*	*JK2*	*Bootstrap1*	*Bootstrap2*	*Bootstrap3*
x1	0.1380	0.2551	0.1209	0.2957	0.2322	0.2679
x2	0.0320	0.0990	0.0318	0.0614	0.0579	-0.0257
x3	0.2228	0.1895	0.1967	0.2656	0.1720	0.2843
x4	0.0519	0.0584	-0.0248	0.0462	0.0364	-0.0636
x5	0.2228	0.1895	0.1967	0.2656	0.1720	0.2843
x6	0.0750	0.0646	-0.0207	0.0688	0.0375	-0.0421
x7	0.2202	0.2171	0.1753	0.2060	0.2540	0.3242
x8	0.0642	0.1213	0.0614	0.0498	0.1912	0.0446
y1	0.2925	0.2401	0.2987	0.2502	0.3220	0.2162
y2	0.2572	0.2053	0.2517	0.2193	0.2954	0.2221
y3	0.2207	0.2126	0.1769	0.2586	0.2033	0.2381
y4	0.2354	0.2420	0.1479	0.2557	0.1770	0.2236
y5	0.1792	0.2260	0.1769	0.2502	0.2264	0.1315
y6	0.0955	0.2625	0.1370	0.2312	0.3277	0.1934
y7	0.3195	0.2741	0.2797	0.3334	0.1514	0.2620
y8	0.3457	0.2634	0.3555	0.1865	0.2405	0.1491
Average (All)	0.1858	0.1950	0.1601	0.2028	0.1936	0.1694
Average (Y)	0.2432	0.2407	0.2280	0.2481	0.2430	0.2045

Table 6: Simulation Results – Relative Standard Error of the Variance Estimator

Variable	*JKn*	*JKn_RG*	*JK2*	*Bootstrap1*	*Bootstrap2*	*Bootstrap3*
x1	0.1382	0.2551	0.0703	0.1949	0.1407	0.1143
x2	0.1787	0.0990	0.0809	0.1698	0.1337	0.0973
x3	0.1344	0.1895	0.0760	0.2013	0.1350	0.1112
x4	0.1822	0.0584	0.0804	0.1712	0.1346	0.0918
x5	0.1344	0.1895	0.0760	0.2013	0.1350	0.1112
x6	0.1759	0.0646	0.0792	0.1749	0.1332	0.0912
x7	0.1411	0.2171	0.0705	0.1778	0.1457	0.1098
x8	0.1694	0.1213	0.0812	0.1665	0.1483	0.1028
y1	0.1367	0.2401	0.0594	0.1853	0.1432	0.0919
y2	0.1331	0.2053	0.0613	0.1873	0.1398	0.0960
y3	0.1384	0.2126	0.0607	0.1939	0.1353	0.0966
y4	0.1489	0.2420	0.0611	0.1956	0.1358	0.0997
y5	0.1747	0.2260	0.1088	0.2090	0.1670	0.1279

Table 6: Simulation Results – Relative Standard Error of the Variance Estimator (Continued)

Variable	*JKn*	*JKn_RG*	*JK2*	*Bootstrap1*	*Bootstrap2*	*Bootstrap3*
y6	0.1827	0.2625	0.1051	0.1962	0.1714	0.1339
y7	0.1364	0.2741	0.0561	0.1925	0.1243	0.1004
y8	0.1363	0.2634	0.0622	0.1786	0.1351	0.0863
Average (All)	0.1526	0.1950	0.0743	0.1872	0.1411	0.1039
Average (Y)	0.1484	0.2407	0.0718	0.1923	0.1440	0.1041

The relative biases are mostly positive. For continuous variables, the bias is negligible when the variable is independent of CVC. However, when it is dependent (correlated) of CVC, the bias is positive and can be as large as 30 percent. For binomial variables, the bias is positive for all variables regardless whether it is correlated or not with CVC and it reaches over 30 percent occasionally. The last two rows in Table 5 provide average of the relative biases. The row labeled as "Average (All)" shows averages of relative biases for all the 16 variables, whereas the row labeled as "Average (Y)" shows averages of relative biases for the 8 binomial variables. The survey contains mainly categorical variables, thus the performance of a variance estimator for the binomial variables is more important. The six variance estimators performed similarly in terms of the relative bias.

However, the variance estimators show big difference in their performance in terms of the coefficient of variation (i.e., relative SE). The paired jackknife (JK2) is the clear winner, and Bootstrap3 is the close second. The random group jackknife (JKn_RG) is the worst. The variance of the bootstrap variance estimator can be reduced by increasing the bootstrap sample size. The bootstrap variance estimator with a bootstrap sample size of 700 or 800 would perform similarly with the JK2 and has an advantage for more complex point estimates such as quantiles. However, since the MTP survey will collect data for mostly categorical variables, the JK2 seems to be the best choice. So we recommend using the JK2 variance estimator for the MTP survey.

Table 7 shows the effect of a positive bias in the variance estimator on the 95 percent confidence interval. We expect about 2 percent over-coverage in average by the JK2 variance estimator.

Table 7: Effect of Positive Bias of the Variance Estimator on the 95% Confidence Interval

Relative bias in variance	*Coverage of 95% confidence interval*	*Over-coverage*
0.00	95.00%	0.00%
0.05	95.54%	0.54%
0.10	96.02%	1.02%
0.15	96.44%	1.44%
0.20	96.82%	1.82%
0.25	97.16%	2.16%
0.30	97.46%	2.46%
0.35	97.72%	2.72%

5. Summary and Discussion

The variance estimation is challenging for the MTP survey due to the special design of

the survey, i.e., only one cancer type is selected for each eligible facility on the frame. To overcome this challenge, this paper studied a few possible variance estimation methods using collapsed strata and tested their performance through a simulation study. The JK2 collapsed variance estimator is the best choice for the MTP survey, though there could be some positive bias in the variance.

There are some suggestions for reducing the positive bias in the collapsed variance estimator in the literature. For example, Wolter (2007) suggests to use the finite population correction of $(1 - 2/N_h)$, where N_h is the population size of cancer types in collapsed stratum h. Shapiro and Bateman (1978) studied the Yates and Grundy (1953) variance estimator assuming Durbin PPS sampling method is used in the collapsed strata. Some rough calculations using the simulated data show that Wolter's simple correction could eliminate the most of the bias for binomial variables. We would like to study this in depth in the future.

References

Rao, J.N.K., and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of American Statistical Association*, 83, 231-241.

Rao, J.N.K.,Wu, C.F.J., and Yue, K. (1992). Some recent work on resampling methods for complex surveys, *Survey Methodology*, 18, 209-217.

Shapiro, G.M., and Bateman, D.V. (1978). A better alternative to the collapsed stratum variance estimate. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 451-456.

Wolter, K.M. (2007). *Introduction to Variance Estimation, 2nd ed*. New York: Springer.

Yates, F., and Grundy, P.M. (1953). Selection without replacement from within strata probability proportional to size. *Journal of Royal Statistical Society B*, 15, 253-261.

Appendix A

Table for the Distribution of the Sampling Frame by Cancer Type and Case Volume

Group number	*HN*	*GI*	*Lung*	*Breast*	*GYN*	*BCNS*	*Count*	*JK2 grouping*
0	0	0	0	0	0	0	17	-
1	0	0	0	0	0	1	8	1
2	0	0	0	1	0	0	1	2
3	0	0	0	1	2	0	1	2
4	0	0	0	2	2	0	1	3
5	0	0	1	0	0	0	1	4
6	0	1	0	0	0	0	4	5
7	0	1	0	1	0	0	12	6
8	0	1	0	1	1	0	1	6
9	0	1	0	2	2	0	1	7
10	0	1	1	0	0	0	4	7
11	0	1	1	0	1	0	1	8
12	0	1	1	0	2	1	1	9
13	0	1	1	1	0	0	145	9
14	0	1	1	1	0	1	19	10

Table for the Distribution of the Sampling Frame by Cancer Type and Case Volume (Continued)

Group number	*HN*	*GI*	*Lung*	*Breast*	*GYN*	*BCNS*	*Count*	*JK2 grouping*
15	0	1	1	1	0	2	1	11
16	0	1	1	1	1	0	196	12
17	0	1	1	1	1	1	83	13
18	0	1	1	1	1	2	3	13
19	0	1	1	1	2	0	5	14
20	0	1	1	1	2	1	6	15
21	0	1	1	1	2	2	1	16
22	0	1	1	2	0	0	1	17
23	0	1	1	2	0	1	2	18
24	0	1	1	2	1	0	9	18
25	0	1	1	2	1	1	18	19
26	0	1	1	2	2	0	3	19
27	0	1	1	2	2	1	3	20
28	0	1	1	2	2	2	1	20
29	0	1	2	1	1	0	5	21
30	0	1	2	1	1	1	8	21
31	0	1	2	1	1	2	3	22
32	0	1	2	1	2	0	1	22
33	0	1	2	2	1	0	1	23
34	0	1	2	2	2	1	1	24
35	0	1	2	2	2	2	1	25
36	0	2	1	1	0	1	1	26
37	0	2	1	1	1	1	5	27
38	0	2	1	1	1	2	1	27
39	0	2	1	1	2	0	1	28
40	0	2	1	2	1	0	1	29
41	0	2	1	2	1	1	2	30
42	0	2	1	2	1	2	1	30
43	0	2	1	2	2	1	2	31
44	0	2	2	1	0	1	1	31
45	0	2	2	1	1	0	2	31
46	0	2	2	1	1	1	2	32
47	0	2	2	1	2	1	2	32
48	0	2	2	1	2	2	1	33
49	0	2	2	2	1	0	2	33
50	0	2	2	2	1	1	3	34
51	0	2	2	2	1	2	1	34
52	0	2	2	2	2	1	5	35
53	0	2	2	2	2	2	2	35
54	1	1	0	1	2	1	1	36
55	1	1	1	1	0	0	19	37
56	1	1	1	1	0	1	7	37
57	1	1	1	1	1	0	77	37
58	1	1	1	1	1	1	86	37

Table for the Distribution of the Sampling Frame by Cancer Type and Case Volume (Continued)

Group number	*HN*	*GI*	*Lung*	*Breast*	*GYN*	*BCNS*	*Count*	*JK2 grouping*
59	1	1	1	1	1	2	1	38
60	1	1	1	1	2	0	1	38
61	1	1	1	1	2	1	12	38
62	1	1	1	2	0	0	1	39
63	1	1	1	2	1	0	3	40
64	1	1	1	2	1	1	24	41
65	1	1	1	2	2	0	2	41
66	1	1	1	2	2	1	4	42
67	1	1	2	1	0	1	1	42
68	1	1	2	1	1	0	11	43
69	1	1	2	1	1	1	39	43
70	1	1	2	1	1	2	1	44
71	1	1	2	1	2	0	1	44
72	1	1	2	1	2	1	2	45
73	1	1	2	1	2	2	1	46
74	1	1	2	2	1	0	1	46
75	1	1	2	2	1	1	17	47
76	1	1	2	2	1	2	1	48
77	1	1	2	2	2	1	2	48
78	1	2	1	1	1	0	1	49
79	1	2	1	1	1	1	3	50
80	1	2	1	1	2	1	2	51
81	1	2	1	2	1	0	1	51
82	1	2	1	2	1	1	5	52
83	1	2	1	2	2	1	7	53
84	1	2	1	2	2	2	2	53
85	1	2	2	1	1	0	3	53
86	1	2	2	1	1	1	10	54
87	1	2	2	1	1	2	6	55
88	1	2	2	1	2	0	1	56
89	1	2	2	1	2	1	4	57
90	1	2	2	1	2	2	4	58
91	1	2	2	2	1	0	1	59
92	1	2	2	2	1	1	45	60
93	1	2	2	2	1	2	8	60
94	1	2	2	2	2	0	2	61
95	1	2	2	2	2	1	48	62
96	1	2	2	2	2	2	45	63
97	2	0	1	1	0	2	1	64
98	2	1	1	1	1	1	3	65
99	2	1	1	1	1	2	2	65
100	2	1	1	1	2	1	2	66
101	2	1	1	1	2	2	1	67
102	2	1	2	1	0	2	1	67

Table for the Distribution of the Sampling Frame by Cancer Type and Case Volume (Continued)

Group number	*HN*	*GI*	*Lung*	*Breast*	*GYN*	*BCNS*	*Count*	*JK2 grouping*
103	2	1	2	1	1	1	2	68
104	2	1	2	1	1	2	2	69
105	2	1	2	1	2	1	1	70
106	2	1	2	2	1	1	2	70
107	2	2	1	1	1	2	2	71
108	2	2	1	1	2	1	4	72
109	2	2	1	1	2	2	2	73
110	2	2	1	2	1	1	1	74
111	2	2	1	2	1	2	1	75
112	2	2	1	2	2	2	2	76
113	2	2	2	1	1	1	3	76
114	2	2	2	1	1	2	3	77
115	2	2	2	1	2	1	1	77
116	2	2	2	1	2	2	2	78
117	2	2	2	2	1	1	10	79
118	2	2	2	2	1	2	12	80
119	2	2	2	2	2	0	1	81
120	2	2	2	2	2	1	27	82
121	2	2	2	2	2	2	104	83
Total							1,309	

Appendix B

Variance Estimation Formulae

B.1 Jackknife Variance Estimator

To define a jackknife variance estimator, we need to define the replicate weights. We first start defining replicate weights for the facilities in the first variance stratum (i.e., $h = 1$) in Table 2. Let g be the index for variance units (either single facilities or random groups) in a variance stratum, and let G_h be the number of variance units for variance stratum h - the G_h are given in the last two columns in Table 2. Suppose that facility i is in the first variance stratum and in the first variance unit (either a single facility or a random group of facilities) (i.e., $g = 1$). Then the r-th replicate weight for the facility if $r \leq G_1$ is defined by:

$$w_i^{(r)} = \begin{cases} 0 & \text{if } i \in g \text{ and } i \in h \\ \{G_h/(G_h - 1)\}\, w_i & \text{if } i \text{ not in } g \text{ but } i \in h \end{cases} \quad \text{(B.1)}$$

For other r's, the replicate weight is the same as the full sample weight w_i, namely,

$$w_i^{(r)} = w_i \quad \text{(B.2)}$$

In general, for r that is $G_1 + G_2 + \cdots + G_{h-1} < r \leq G_1 + G_2 + \cdots + G_h$, the replicate weight is defined by (B.1), and for other r's by (B.2). In this way, a replicate is represented by their replicate weights.

Then the jackknife variance estimator for the sample estimate $\hat{\theta}$ for a population parameter θ is given by:

$$\hat{V}(\hat{\theta}) = \sum_{r=1}^{R} c_r \left(\hat{\theta}^{(r)} - \hat{\theta}\right)^2 \quad \text{(B.3)}$$

where $c_r = (G_h - 1)/G_h$ for $G_1 + G_2 + \cdots + G_{h-1} < r \leq G_1 + G_2 + \cdots + G_h$, and $\hat{\theta}^{(r)}$ is the r-th replicate estimate for θ, based on the r-th replicate weights. The same formula to calculate $\hat{\theta}$ is also used to calculate the replicate estimate but the replicate weights are used instead of the full sample weights. Note that c_r is the same for all replicates defined in the same variance stratum.

The total number of replicates, R, for the full jackknife (JKn) is 1,278, and that for the random group jackknife (JKn_RG) is 210 after removing absolute certainties in replicate formation.

The paired jackknife defines the replicate by dropping one variance unit randomly from each variance stratum, resulting in only one replicate for each variance stratum. The replicate weight for the retained variance unit is then assigned 2 times of the full sample weight, whereas the dropped one is given a zero weight. Then the variance estimate is obtained from (B.3) with $c_r = 1$.

B.2 Bootstrap Variance Estimator

We used the rescaled bootstrap method originally proposed by Rao and Wu (1988) and later modified by Rao, Wu, and Yue (1992). The simpler form of the estimator defines the b-th bootstrap weight for facility i in variance stratum h by:

$$w_i^{(b)} = w_i \, t_i^{(b)} n_h/(n_h - 1) \quad \text{(B.4)}$$

where $t_i^{(b)}$ is the number of times facility i is selected in the b-th bootstrap sample, and n_h is the number of facilities in variance stratum h. Note that $t_i^{(b)}$ can be zero if facility i is not selected at all in the b-th bootstrap sample. The bootstrap variance estimator is given by

$$\hat{V}(\hat{\theta}) = \sum_{b=1}^{B} \left(\hat{\theta}^{(b)} - \hat{\theta}\right)^2 / B \quad \text{(B.5)}$$

where B is the bootstrap sample size, and $\hat{\theta}^{(b)}$ is calculated using the bootstrap weight.